Project Report

Tongliang Deng
SIST
ShanghaiTech University
Shanghai, China
dengtl@shanghaitech.edu.cn

Abstract

In this project, we reimplemented the idea of paper *Designing Statistical Estimators That Balance Sample Size, Risk, and Computational Cost* proposed to tradeoff between cost and risk. And, we found there is a bug of the presudocode given by this paper, and presented the corrected version in this report.

1 Introduction

As the data increase, we have to use more efficient algorithm to handle this massive data. In common expectation, computational resource should be increased as the data increasing. The amount of computational power available, however, is growing slowly relative to sample sizes. Thus, large-scale problems need more time to solve. This creates a demand for new algorithms that offer better performance when presented with large data sets.

1.1 Regularized linear regression

Assume there is data set $\{(\boldsymbol{a_i}, b_i) : i = 1, \cdots, m\}$ with m *samples*, where $\boldsymbol{a_i} \in \mathbb{R}^d$ are *inputs*, and the $b_i \in \mathbb{R}$ are the *responses* of a statistical model. Call m as the *sample size*, and in our experiment where $m < d$. Given a vector of parameter $\boldsymbol{x}^\natural \in \mathbb{R}^d$, relate the inputs and responses through the linear model

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}^\natural + \boldsymbol{v}, \boldsymbol{b} \in \mathbb{R}^m, \boldsymbol{A} \in \mathbb{R}^{m \times d}$$

where the ith row of $\boldsymbol{A}$ as the input $\boldsymbol{a}_i$, the ith entry of $\boldsymbol{b}$ is b_i, and the entries of $\boldsymbol{v} \in \mathbb{R}^m$ are independent, zero-mean random variates. The goal of the regression problem is to infer the underlying parameters $\boldsymbol{x}^\natural$ from the data matrix $\boldsymbol{A}$.

Denote $\hat{\boldsymbol{x}}$ as an estimate of the true vector $\boldsymbol{x}^\natural$. Then the *average squared prediction error* of $\hat{\boldsymbol{x}}$ is

$$R(\hat{\boldsymbol{x}}) = \frac{1}{m}||\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{x}^\natural||^2.$$

which is called *statistical risk*. However, in most cases, we don't have true vector $\boldsymbol{x}^\natural$. We could have (noisy) observations $\boldsymbol{b}$ in the given data set, then compute

$$\hat{R}(\hat{\boldsymbol{x}}) = \frac{1}{m}||\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{b}||^2.$$

The quantity $\hat{R}(\hat{\boldsymbol{x}})$ is an estimation of $R(\hat{\boldsymbol{x}})$., which is called *empirical risk*.

In liner regression model, the goal is minimizing empirical risk. Since our problem has fewer samples than the number of parameters ($m < d$), then we solve the regularized linear regression problem

$$\hat{x} := \arg\min_x f(x), \text{ subject to } ||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}|| \leq \sqrt{m \cdot R_{max}} =: \epsilon$$

where the proper convex function f is a *regularizer*, and R_{max} is the maximal empirical risk we can tolerate.

2 The geometry of the time-data tradeoff

At geometry level, there is a opportunity for us to do the time-data tradeoff.

2.1 Descent Cones and Statistical Dimension

The descent cone of a proper convex function f at the point $\boldsymbol{x} \in \mathbb{R}^d$ is the convex cone

$$\mathcal{D}(f; x) := \bigcup_{\tau > 0} \{\boldsymbol{y} \in \mathbb{R}^d : f(\boldsymbol{x} + \tau \boldsymbol{y}) \leq f(\boldsymbol{x})\}$$

And, *statistical dimension* is used to quantify the size of this convex cone, it defined as

$$\delta(\mathcal{C}) := \mathbb{E}_g[||\Pi_{\mathcal{C}}(g)||^2]$$

where $\boldsymbol{g} \in \mathbb{R}^d$ has independent standard Gaussian entries, and $\Pi_{\mathcal{C}}$ is the projection operator onto $\mathcal{C}$. In this experiment, $\delta(\mathcal{D}(f; \boldsymbol{x}^\natural))$ plays the critical role.

Amelunxen et al. [1] proved that, under certain randomized data models with noiseless measurements, the regularized linear regression problem undergoes a phase transition when the number m of samples equals $\delta(\mathcal{D}(f; \boldsymbol{x}^\natural))$. Oymak and Hassibi [2] characterized the stability of this phase transition in the presence of noise.

The phase transition gives us that, when the number m is smaller than $\delta(\mathcal{D}(f; \boldsymbol{x}^\natural))$, the regularized linear regression problem has no robustness to noise. That is, as the number of samples increases towards the phase transition, the statistical accuracy of the solution does not improve. After crossing the phase transition point, however, additional samples decrease the worst-case risk at the rate m^{-1}.

In our experiments, we set R_{max} and ϵ as

$$R_{max} = \sigma^2 \Big(1 - \frac{\delta(\mathcal{D}(f; \boldsymbol{x}^\natural))}{m}\Big)$$

$$\epsilon = \sigma(m - \delta(\mathcal{D}(f; \boldsymbol{x}^\natural)))$$

Chandrasekaran and Jordan [3] shows that enlarging convex constraint sets can make the optimization problems easier to solve. However, it creates a loss of statistical accuracy.

As Figure 1 shows that by enlarging the sublevel sets of the regularizer f, which increase the statistical dimension of the descent cone of f at $\boldsymbol{x}^\natural$. The relaxed regularizer results in a problem that is easier to solve computationally, then there is a tradeoff between sample size, computational time, and statistical accuracy.

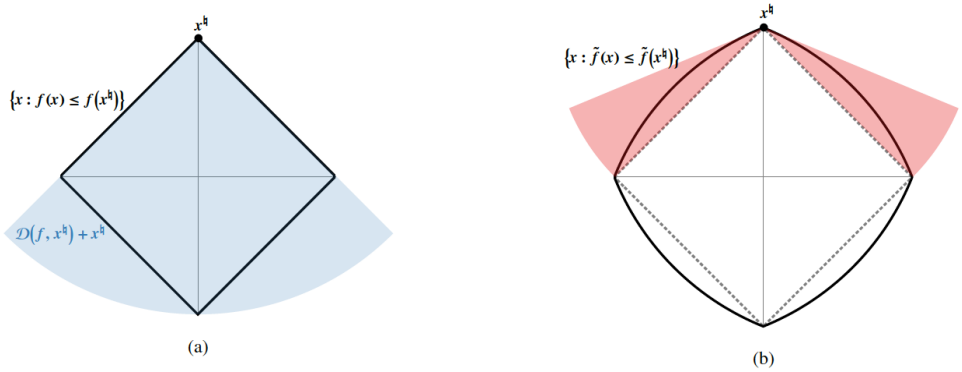

Figure 1: A geometric opportunity. Panel (a) illustrates the sublevel set and descent cone of a regularizer f at the point $\boldsymbol{x}^\natural$. Panel (b) shows a relaxed regularizer $\tilde{f}$ with larger sublevel sets. The shaded area indicates the difference between the descent cones of $\tilde{f}$ and f at $\boldsymbol{x}^\natural$.

3 A time-data tradeoff via dual-smoothing

At computational level, we could confirm that there is computational benefit of smoothing optimization problems. In our experiment, we mainly focus on smoothing dual problem of the primal problem.

Say a function $f_\mu : \mathbb{R}^d \to \mathbb{R}$ is *μ-strongly convex* if there exist a positive constant μ such that the the following function is convex.

$$\boldsymbol{x} \to f_\mu(\boldsymbol{x}) - \frac{\mu}{2}||\boldsymbol{x}||^2.$$

Say a function $g : \mathbb{R}^d \to \mathbb{R}$ has a *L-Lipschitz gradient* if there exists a positive constant such that for any $z_1, z_2 \in \mathbb{R}^d$ we have

$$||\nabla g(z_1) - \nabla g(z_2)|| \leq L||z_1 - z_2||.$$

Nemirovski and Yudin [4] show that the best achievable convergence rate that minimizes a convex objective with a Lipschitz gradient is $O(1/\gamma^2)$ iterations, where γ is the numerical accuracy.

If smooth f by any amount, its descent cones become halfspaces, and we lose all control over their size. We instead consider a method to smooth the dual of the optimization problem. This technique preserves the geometric opportunity while allowing for a computational speedup.

The primal problem is

$$\hat{x}_\mu := \arg\min_x f_\mu(x), \text{ subject to } ||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}|| \leq \epsilon.$$

where f_μ is the replacer of f, and f_μ is *μ-strongly* convex function.

The dual problem is then

$$\text{maximize } g_\mu(\boldsymbol{z}) := \inf_x \left\{ f_\mu(\boldsymbol{x}) - \langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \rangle - \epsilon||\boldsymbol{z}|| \right\}$$

Although g_μ is not smooth, rewrite it to the following form

$$\begin{aligned} g_\mu(\boldsymbol{z}) &= \inf_x \left\{ f_\mu(\boldsymbol{x}) - \langle \boldsymbol{A^z}, \boldsymbol{x} \rangle - \epsilon||\boldsymbol{z}|| \right\} \\ &= -f_\mu^\star(\boldsymbol{A^T z}) + \langle \boldsymbol{z}, \boldsymbol{b} \rangle - \epsilon||z|| \end{aligned} \tag{1}$$

where $f_\mu^\star$ is the convex conjugate of f_μ. we have that $\tilde{g}_\mu(z) = -f_\mu^\star(\boldsymbol{A^T z}) + \langle \boldsymbol{z}, \boldsymbol{b} \rangle$ has a Lipschitz gradient $L_\mu := \mu^{-1} \|A\|^2$. And, $h(z) = \epsilon \|z\|$ is nonsmooth.

Then, we can now solve the composite dual problem (1) using an accelerated gradient method. In the following give the Auslender–Teboulle algorithm as example, and our experiment use this algorithm to solve the dual problem.

Algorithm 1: Auslender-Teboulle

Input : measurement matrix $\boldsymbol{A}$, observed vector $\boldsymbol{b}$, parameter ϵ

1 $z_0 \leftarrow \boldsymbol{0}, \bar{z}_0 \leftarrow z_0, \theta_0 \leftarrow 1$;
2 **for** *k = 0,1,2,...* **do**
3 $\quad \boldsymbol{y_k} \leftarrow (1 - \theta_k) z_k + \theta_k \bar{z}_k$;
4 $\quad \boldsymbol{x_k} \leftarrow \arg\min_x \{ f(\boldsymbol{x}) + \langle \boldsymbol{y_k}, \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \rangle \}$;
5 $\quad \bar{z}_{k+1} \leftarrow \text{Shrink}(\bar{z}_k + (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x_k})/(L_\mu \cdot \theta), \epsilon/(L_\mu \cdot \theta))$;
6 $\quad z_{k+1} \leftarrow (1 - \theta_k) z_k + \theta_k \bar{z}_{k+1}$;
7 $\quad \theta_{k+1} \leftarrow 2/(1 + (1 + 4/\theta_k^2)^{1/2})$;
8 **end**

In Algorithm 1, the line 5 would be implemented as

$$\text{Shrink}(\boldsymbol{z}, t) = \max\{1 - \frac{t}{||\boldsymbol{z}||}, 0\} \cdot \boldsymbol{z}$$

Since the regularizer f_μ is μ-strongly convex. Apply Algorithm 1 to the corresponding dual problem (1), and let $\boldsymbol{z}^\star$ be the optimal dual point, then at the k^{th} iteration, we have

$$|\,\|\boldsymbol{A}\boldsymbol{x_k}-\boldsymbol{b}\|-\epsilon| \leq \frac{2\,\|\boldsymbol{A}\|^2\,\|\boldsymbol{z}^\star\|}{\mu k}$$

This bound suggests that, as the convexity of the regularizer f_μ increases, the number of iterations sufficient for Algorithm 1 to converge to the preset empirical risk target R_{max} decreases.

4 Smoothing schemes

Many common regularizers such as the l_1 norm are not strongly convex, and so we must provide an appropriate relaxation method before applying Algorithm 1.

Introduce a family $\{f_\mu : \mu > 0\}$ of strongly convex majorants:

$$f_\mu(\boldsymbol{x}) := f(\boldsymbol{x}) + \frac{\mu}{2}||\boldsymbol{x}||^2$$

Clearly, f_μ is μ-strongly convex, and they are relaxations which allow us to realize the tradeoff.

4.1 Choosing a smoothing parameter

In our experiments, we did three type of smoothing scheme, *constant smoothing*, *constant risk* and *tunable balance*.

- *Constant Smoothing*: Larger values of μ lead to larger values of $\delta(\mathcal{D}(f_\mu;\boldsymbol{x}))$, the location of the phase transition. And it lead to higher worst-case level of statistical-risk. In this case we choose $\bar{\mu} = 0.1$, and calculate its corresponding $\bar{\delta}$, and set baseline number of samples $\bar{m} = \bar{\delta} + \sqrt{d}$. Set $\bar{m}$ in this way, since we need to be sure that m bigger than the transition phase.
- *Constant Risk*: In different m, set their smoothing parameter μ, such that

$$\frac{\delta(\mathcal{D}(f_\mu;\boldsymbol{x}^\natural))}{m} = \frac{\bar{\delta}}{\bar{m}}$$

- *A Tunable Balance*: To get a balance between the above two cases, we could choose a balancing parameter α, such that

$$\frac{\delta(\mathcal{D}(f_\mu;\boldsymbol{x}^\natural)}{m} = \frac{\bar{\delta}}{\bar{m}+(m-\bar{m})^\alpha}$$

5 Experiments

In this section, we did *sparse vector regression* and *low-rank matrix regression* from the paper. Since our computer's computing resource is not enough to do the image part of the paper, the dimension of this part is too large to do it.

5.1 Sparse vector regression

In our experiment, the parameter vector $\boldsymbol{x}^\natural \in \mathbb{R}^d$ in the data model is sparse, and choose l_1 norm as the regularizer. Then appy the dual-smoothing procedure to obtain the relaxed regularizer

$$f_\mu(\boldsymbol{x}) = \|\boldsymbol{x}\|_{l_1} + \frac{\mu}{2}\|\boldsymbol{x}\|^2$$

To apply Algorithm 1 to the dual-smoothed sparse vector regression problem, we implemented the line 4 of Algorithm 1 as

$$\boldsymbol{x}_k \leftarrow \mu \cdot \text{SoftThresh}(\boldsymbol{A^T y_k}, 1)$$
$$[\text{SoftThresh}(\boldsymbol{x}, t)]_i = \text{sign}(x_i)\cdot\max\{|x_i|-t, 0\}$$

To choose the smoothing parameter μ, we have to calculate the statistical dimension δ. To calculate it in the l_1 norm scheme. We firstly define the normalized sparsity $\rho := s/d$, then

$$\frac{\delta(\mathcal{D}(f_\mu; \boldsymbol{x}))}{d} \leq \psi(\rho)$$

where $\psi : [0, 1] \to \mathbb{R}$ is

$$\psi(\rho) = \inf_{\tau \geq 0} \left\{ \rho[1 + \tau^2(1 + \mu||\boldsymbol{x}||_\infty)^2] + (1 - \rho)\sqrt{\frac{2}{\pi}} \int_\tau^\infty (\mu - \tau)^2 e^{-u^2/2} \mathrm{d}u \right\}$$

5.1.1 Numerical experiment

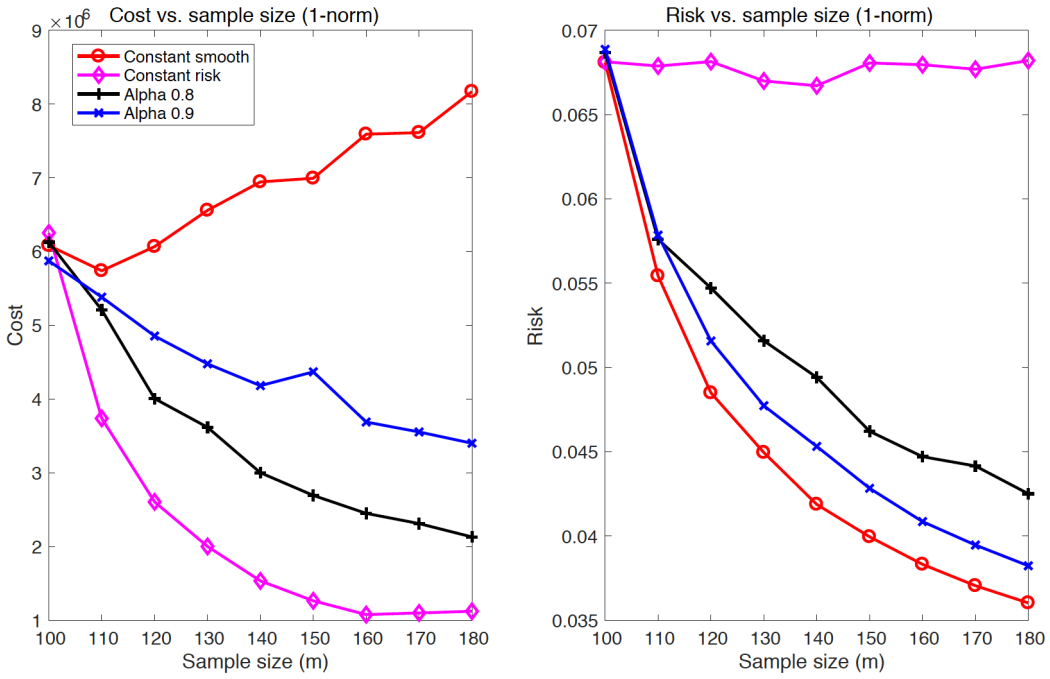

Figure 2: This shows 200 random trials of the dual-smoothed sparse vector regression problem with ambient dimension $d = 400$, normalized sparsity $\rho = 5\%$, and noise level $\sigma = 0.1$.

In Figure 2, we show the results of a numerical experiment that reveals the time–data tradeoff enabled by the smoothing schemes.

For the constant smoothing case, we choose $\mu = 0.1$. We compare this to the constant risk and balanced method with $\alpha = 0.9$ and $\alpha = 0.8$.

In this experiment, fix both the ambient dimension $d = 400$ and the normalized sparsity $\rho = 5\%$. In each smoothing approach, we generate and solve 200 random sparse vector regression problems for each value of the sample size $m = 100, 110, 120, \cdots, 180$. Each problem comprises a random measurement matrix $\boldsymbol{A} \in \mathbb{R}^{m \times d}$ with orthonormal rows and a random sparse vector $\boldsymbol{x}^\natural$ whose 20 nonzero entries are ± 1. Baseline smoothing $\bar{\mu} = 0.1$ and $\bar{m} = 100 \approx \delta(\mathcal{D}(f_{\bar{u}}; \boldsymbol{x}^\natural)) + \sqrt{400}$. And set the stop criterion as that when $|\,\|\boldsymbol{A}\boldsymbol{x_k} - \boldsymbol{b}\| - \epsilon| < 10^{-3}$.

5.2 Low-rank matrix regression

Let $\boldsymbol{X}^\natural \in \mathbb{R}^{d_1 \times d_2}$ be the true matrix, and let $\boldsymbol{A} \in \mathbb{R}^{m \times d}$ be a measurement matrix, where $d = d_1 d_2$. Then the observations are given by $\boldsymbol{b} = \boldsymbol{A} \cdot vec(\boldsymbol{X}^\natural)$, where vec returns the (column) vector obtained by stacking the columns of the input matrix. And, $\boldsymbol{X}^\natural$ is low-rank. The regularizer is $f = ||\cdot||_{S_1}$ which is Schatten 1-norm

$$f_\mu(\boldsymbol{X}) = \|X\|_{S_1} + \frac{\mu}{2}\|\boldsymbol{X}\|_F^2$$

To apply Algorithm 1 to the dual-smoothed sparse vector regression problem, we implemented the line 4 of Algorithm 1 as

$$\boldsymbol{X_k} \leftarrow \mu \cdot \text{SoftThreshSingVal}(\text{mat}(\boldsymbol{A^T y_k}), 1)$$

In *SoftThreshSingVal* we do singular decomposition on the matrix $\boldsymbol{A^T y_k}$, in the following form

$$\boldsymbol{X} = \boldsymbol{U} \cdot \text{diag}(\boldsymbol{\sigma}) \cdot \boldsymbol{V}^T$$

$$\text{SoftThreshSingVal}(\boldsymbol{X}, t) = \boldsymbol{U} \cdot \text{diag}(\text{SoftThresh}(\boldsymbol{\sigma}, t)) \cdot \boldsymbol{V}^T$$

where *SoftThresh* is the same as sparse vector regression part.

To choose the smoothing parameter μ, we have to calculate it. Let $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$ have rank r, and define the normalized rank $\rho = r/d_1$, then

$$\frac{\delta(\mathcal{D}(f_\mu; \boldsymbol{X}))}{d_1^2} \leq \psi(\rho) + o(1)$$

where $\psi : [0, 1] \to \mathbb{R}$ is

$$\psi(\rho) := \inf_{0 \leq \tau \leq 2} \Big\{ \rho + (1-\rho) \Big[\rho(1 + \tau^2(1 + \mu \|\boldsymbol{X}\|)^2) + \frac{(1-\rho)}{12\pi} [24(1+\tau^2)\cos^{-1}(\tau/2) - \tau(26+\tau^2)\sqrt{4-\tau^2}] \Big] \Big\}$$

5.2.1 Numerical experiment

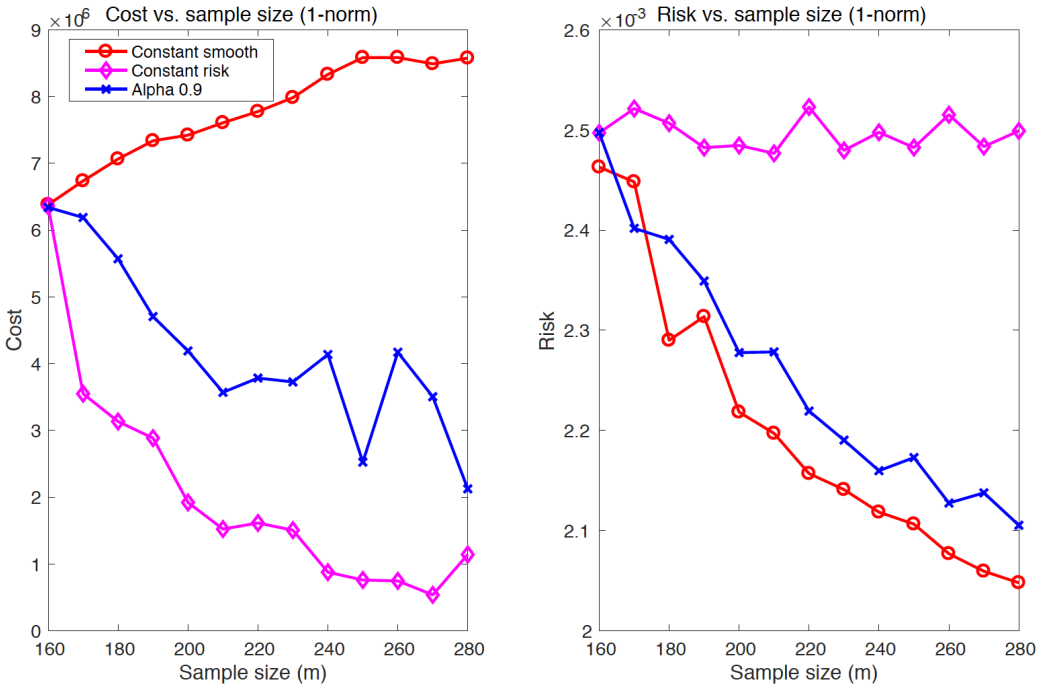

Figure 3: This shows 200 random trials of the dual-smoothed sparse vector regression problem with ambient dimension $d_1 \times d_2 = 20 \times 20$, normalized rank $\rho = 5\%$, and noise level $\sigma = 0.1$.

Figure 3 shows the results of a substantially similar numerical experiment to the one performed for sparse vector regression. In our experiment, choose the baseline smoothing parameter $\bar{\mu} = 0.1$. As before, we compare the constant smoothing, constant risk, and balanced $\alpha = 0.9$ schemes.

In this case, we use the ambient dimension $d = 20 \times 20$ and set the normalized rank $\rho = 5\%$. We test each method with 200 random trails of the low-rank matrix regression problem for each value of the sample size $m = 160, 170, 180, \cdots, 280$. Baseline smoothing $\bar{\mu} = 0.1$ and $\bar{m} = 160 \approx \delta(\mathcal{D}(f_{\bar{u}}; \boldsymbol{x}^\natural)) + \sqrt{200 \cdot 200}$. And set the stop criterion as that when $|\,\|\boldsymbol{A}\boldsymbol{x_k} - \boldsymbol{b}\| - \epsilon| < 10^{-3}$.

References

[1] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: A geometric theory of phase transitions in convex optimization," Information and Inference, vol. to appear, 2014.

[2] S. Oymak and B. Hassibi, "Sharp MSE Bounds for Proximal Denoising," arXiv, 2013, 1305.2714v5.

[3] V. Chandrasekaran and M. I. Jordan, "Computational and statistical tradeoffs via convex relaxation," Proc. Natl. Acad. Sci. USA, vol. 110, no. 13, pp. E1181–E1190, 2013.

[4] A. S. Nemirovsky and D. B. Yudin, Problem complexity and method efficiency in optimization. A Wiley-Interscience Publication, New York: John Wiley & Sons Inc., 1983.